第 37 卷第 1 期 计算机应用研究 Vol. 37 No. 1
录用定稿 Application Research of Computers Accepted Paper

基于深度双向模型和特征融合的视频转文字研究 *

宁培阳，史景伦，张荣锋，邱 威

(华南理工大学 电子与信息学院，广州 510640)

摘 要：自动生成视频的自然语言描述，是一个非常具有挑战性的研究热点。基于深度 BLSTM 模型和 CNNs 特征的方法，能够学习到视频序列的全局时空关联信息。针对视频转文字时面临的准确率低以及计算复杂度高的问题，提出了深度 BMGU 模型，从而在保持深度 BLSTM 模型结构优势的同时提高计算效率；还将原始视频帧的 CNN 特征，与经过 Haar 特征预处理后的视频的 CNNs 特征进行后期融合，从而增加了训练特征的多样性，进而提升了视频转自然语言的实验效果。在 M-VAD 和 MPII-MD 数据集中，相对原 S2VT 模型，所提方法分别将 METEOR 分数从 6.7 及 7.1 提高到 8.0 和 8.3。结果表明所提方法有效地改善了原 S2VT 模型的准确率和语言描述效果。

关键词：视频转文字；深度双向模型；哈尔特征；特征融合；卷积神经网络

中图分类号：TP391.41 doi: 10.19734/j.issn.1001-3695.2018.03.0488

Research on video description based on deep bidirectional model and feature fusion

Ning Peiyang, Shi Jinglun, Zhang Rongfeng, Qiu Wei

(*School of Electronic & Information Engineering*, *South China University of Technology*, *Guangzhou* 510641, *China*)

Abstract: Automatically generating a natural language description of a video is a challenging work for computer vision. The method based on deep bidirectional long-short term memory (DBLSTM) and CNN feature, had the ability to learn global spatiotemporal correlation information of videos. Focusing on the low accuracy and high computational complexity of video to text, this paper proposed a new method, which based on the deep bidirectional minimal gated unit (BMGU) in order to improve the computational efficiency while maintaining the advantages in structure of the deep BLSTM model. In the same time, by merging the CNNs feature of the original frames and the CNNs feature of the frames with Haar feature increased the diversity of training features and improved the effect of the video to text. By using the datasets of M-VAD and MPII-MD, comparing to the original S2VT model, the proposed method is able to increase the scores from 6.7 to 8.0 and from 7.1 to 8.3 in METEOR. The results show that the proposed method can effectively improve the accuracy and the description of the videos of the original S2VT model.

Key words: video to text; deep bidirectional model; haar feature; feature fusion; convolutional neural networks

0 引言

视频转自然语言（video captioning,又称自动生成视频的自然语言描述），其主要任务是对视频进行理解和分析，并进一步获取有用的语义信息，然后，将这些视频帧中的语义信息与应用的语义环境进行关联，从而将视频帧序列转换为自然语言描述[1]。视频转自然语言可用于智能安防、人机交互、视频检索等诸多领域，具有较高的应用价值和现实意义。

随着深度学习在计算机视觉的诸多领域的逐步延伸，以 S2VT[2]（sequence to sequence-video to text）为代表的视频转文字方法，在性能上显著地超越了以往的非深度学习方法，但也仍存在若干方面需要改进。例如，为了获取视频帧中所包含的语义信息，一般先使用 CNN 模型来提取视频帧的卷积特征[3]，卷积特征中包含视频帧的空间信息。然而，视频描述数据集中的视频帧常常存在背景繁杂（存在多种对象）的情况，某些 CNNs 模型提取这类视频帧的特征时性能会降低，导致视频转文字方法不能输出较为准确的自然语言描述。另外，LSTM 是 S2VT 方法的核心模型，它通过将 RNN（recurrent neural networks）无门的结构改进为具有三个门结构和两个隐藏状态的结构，较好地克服了梯度弥散或梯度爆炸的问题，从而具有对长序列信

收稿日期：2018-03-19；**修回日期：**2018-06-04 **基金项目：**国家自然科学基金资助项目（61671213）；广州市人体数据科学重点实验室项目基金资助项目（201605030011）

作者简介：宁培阳（1992-），男，广西南宁人，硕士研究生，主要研究方向为视频理解（540439329@qq.com）；史景伦（1977-），男，教授，博士，主要研究方向为异构传感器与多数据融合、深度学习相关智能算法、智能控制；张荣锋（1980-），男，博士，讲师，主要研究方向为机器学习与视频处理；邱威（1994-），男，硕士研究生，主要研究方向为机器学习与推荐系统.

息进行较好地学习和建模的能力[4, 5]。然而 LSTM 增加了大量的参数，降低了方法的计算效率，不利于将其应用于实时性要求高、计算条件严苛的场合。并且，近年来 Chung 等人[6]通过实验发现，门结构的数量越多并不意味着最终的实验效果会更好，一些较为简单的 RNN 模型在降低了计算复杂度的同时，甚至还能够收到比 LSTM 更好的效果。

针对 S2VT 方法中存在的描述准确率低、计算复杂等问题，本文提出了基于深度双向循环神经网络[7]和哈尔特征[8]（Haar feature）的视频转文字方法，具体如下：首先，针对 S2VT 模型（其编码层基于单向 LSTM）不能充分学习视频序列中前后帧的时序信息的问题，提出基于深度双向 LSTM 的视频转文字方法以学习到全局的时间关联信息。其次，针对视频帧具有背景繁杂的特点，而影响对主体对象的特征提取的问题，提出基于 Haar 特征预处理的视频帧增强方法，即在使用 VGG 等卷积神经网络提取视频帧的隐式特征前，通过提取 Haar 特征对视频帧进行预处理，以达到抑制繁杂背景信息和强化主体对象信息的目的；再次，针对深度 BLSTM 计算复杂度高的问题，提出基于深度 BMGU 的视频转文字方法。实验表明，这种基于简化模型的方法，不仅能够有效地提高计算效率，而且自然语言描述的效果也与深度 BLSTM 模型相当。

1　视频转自然语言原理与 S2VT 模型

视频转自然语言的任务，在数学上可以表述为：给定视频的帧序列 $\mathbf{X}(x_1,x_2,\ldots,x_t,\ldots x_n)$，给出关于概括视频语意信息的词序列 $\mathbf{Y}(y_1,y_2,\ldots,y_t,\ldots y_m)$的条件概率，即

$$p(\mathbf{Y}/\mathbf{X}) = p(y_1,\ldots,y_m|x_1,\ldots,x_n) \tag{1}$$

其中：帧序列长度 n 和词序列长度 m 是可变的。一般地，$n\neq m$ 且 $n>m$。基于循环神经网络的视频转文字方法，通过构造“编码器-解码器（encoder-decoder）”模型，从而使用隐式特征来实现帧序列和词序列的联合建模。相应地，可把式(1)改写为

$$p(\mathbf{Y}/\mathbf{X}) = p(y_1,\ldots,y_m|x_1,\ldots,x_n)=\prod_{t=1}^{m} p(y_t\,|\,h_{n+t-1},y_{t-1}) \tag{2}$$

S2VT[2]具有理解和描述视频内容的功能，它是一个经典的、基于 LSTM 的视频转文字模型，能够生成自然语言的句子来描述视频中所发生的相应事件。

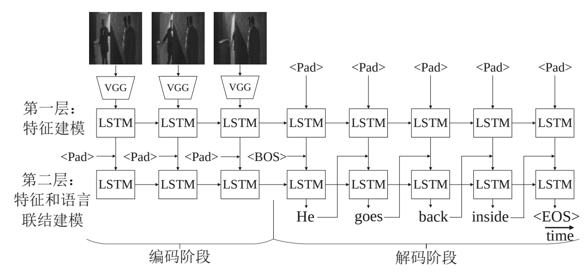


图 1　S2VT 模型原理图

Fig.1　Schematic diagram of S2VT model

如图 1 所示，S2VT 模型通过 VGG-16[9]网络获取输入视频序列的卷积特征（CNNs 特征），再将特征序列按时序地输入第一层 LSTM 进行特征建模；在第二层 LSTM 中，通过 LSTM 网络学习帧序列与词序列之间的映射关系，完成特征和语言的关联建模。另外，图中“<Pad>”表示用全零向量作为输入来填充相应的位信息，输入“<BOS>” 则表明帧序列已输入完毕，用于指示模型从编码阶段切换到解码阶段（即开始预测词序列）。“<EOS>”表示 S2VT 模型预测的词序列已输出完毕。

LSTM 是 S2VT 模型实现式(2)功能的核心算法，具体计算上，假设在时刻 t 输入的变量是 x_t，对应的隐藏层状态参数是 h_t，而记忆单元的状态是 c_t，则在 t 时刻 LSTM 单元中的公式如下[5]：

$$i_t = \sigma(W_{xi}x_t + W_{hi}h_{t-1} + b_i) \tag{3}$$

$$f_t = \sigma(W_{xf}x_t + W_{hf}h_{t-1} + b_f) \tag{4}$$

$$o_t = \sigma(W_{xo}x_t + W_{ho}h_{t-1} + b_o) \tag{5}$$

$$g_t = \varphi(W_{xg}x_t + W_{hg}h_{t-1} + b_g) \tag{6}$$

$$c_t = f_t \odot c_{t-1} + i_t \odot g_t \tag{7}$$

$$h_t = f_t \odot \phi(c_t) \tag{8}$$

在式（3）~（8）中，i、f、o、g 分别表示 LSTM 的输入门、遗忘门、输出门、输入调制栅,对应的各门偏置向量为 b_i、b_f、b_c、b_g。$h_t\in R^n$ 表示 n 个隐藏状态参数。W_{ab}（$a\in\{x,h\}$，$b\in\{i,h,o,g\}$）表示输入或隐藏层状态参数 a 到门 b 的权重矩阵。$\sigma(x)$是 sigmoid 函数，$\varphi(x)$是双曲正切函数，而⊙是逐元素点积（element-wise product）运算。通过式（3）~（8），S2VT 模型依次迭代求出各时刻的隐藏层参数 $h_1,h_2,\ldots h_t,\ldots h_n$，并进一步求出隐藏层参数关于词 y_t（t=1,2,…,m）的条件概率 $p(y_t|h_{n+t})$，从而得到预测的词序列。

2　改进方法的提出

2.1　基于 DBLSTM 与 Haar 特征预处理的视频转文字方法

首先，针对 S2VT 模型的基于单向 LSTM 编码层对视频帧特征利用不充分的问题，采用深度双向 LSTM（DBLSTM）网络对方法进行改进。基于深度双向 LSTM 的视频转文字方法原理图如图 2 所示。

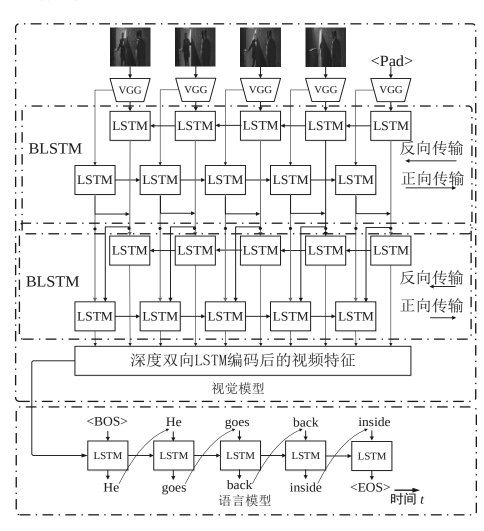


图 2　基于深度双向 LSTM 的视频转文字方法原理图

Fig.2　Schematic diagram of video-to-text based on depth bidirectional LSTM

其中，一个正向传输信息的 LSTM 和一个反向传输信息的 LSTM 组成了 BLSTM，2 个 BLSTM 再按图 3 中的连接组成 DBLSTM。关于两层 LSTM 隐藏层状态参数的融合，一般使用如下的方法[7]：

$$\vec{h}_t = H\left(W_{x\vec{h}}x_t + W_{\vec{h}\vec{h}}\vec{h}_{t-1} + b_{\vec{h}}\right) \tag{9}$$

$$\overleftarrow{h}_t = H\left(W_{x\overleftarrow{h}}x_t + W_{\overleftarrow{h}\overleftarrow{h}}\overleftarrow{h}_{t+1} + b_{\overleftarrow{h}}\right) \tag{10}$$

$$y_t = W_{\vec{h}y}\vec{h}_t + W_{\overleftarrow{h}y}\overleftarrow{h}_t + b_h \tag{11}$$

其中：$\vec{h}_t$、$\overleftarrow{h}_t$ 分别是正向传输层、反向传输层中 LSTM 在 t 时刻的隐藏层状态参数，H(x)为 LSTM 的激活函数，y_t 是融合后的输出。其余参数的含义与式(3)~(8)的参数含义类似。

通过将两个 BLSTM 网络的堆叠而构成深度 BLSTM 网络，来作为深度 RNN 网络的一种改进方法。这使得新网络结构具有下面的优点：首先，BLSTM 除了能够学习到前面帧的相关信息（information in previous frames），还能够学习到未来帧的相关信息（information in future frames），通过前后帧的关联学习以及上下文的关联学习，因而可利用视频中的全局时间信息以增强视频-句子对的学习效果，从而提高了视频转自然语言的准确率。这样就可以有效克服单向 LSTM 只能利用前面帧的相关信息的局限性。其次，在深度神经网络（deep neural network）中，通过拓宽网络的宽度（wider and wider）和增加网络的深度（deeper and deeper）是优化并提高模型性能的两个主要方向：相对应于 CNNs 网络在空间上的深度，LSTM 则是时间上的深度网络，双向 LSTM（BLSTM）相对单向 LSTM 是在时间维度上更深(deeper in time)的网络，因此，BLSTM 相比 LSTM 增强了网络在时间上的依赖性，而深度 BLSTM 则进一步强化了其时间上的依赖性。增加时间深度的方式，增加了网络的参数，也使得训练时可以增强视频和自然语言的关联学习，从而进一步提高了视频转文字的训练和学习效果，进而提高了视频转文字的实验效果（在实验结果上的直观表现为 METEOR 分数的提高）。当然，更深的网络就包含了更多的参数，因而也增加了计算复杂度。

其次，针对视频繁杂的背景可能影响 CNN 对主体特征的提取的情况，对视频帧按 RGB 三个通道进行分离，并逐通道进行一阶 Haar 小波滤波，滤除细节信息，最后重组，从而得到包含了 Haar 特征的视频帧。流程如图 3 所示。

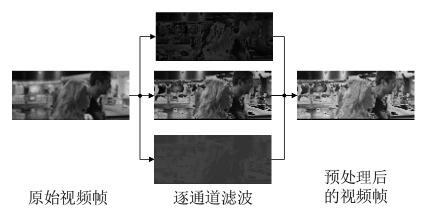


图 3　视频帧提取 Harr 特征的流程图

Fig.3　Flowchart of extracting Harr features fron video frames

由对比图 3 中处理前后的图片可知，经过提取 Haar 特征后的视频帧中，视频帧的主体信息得到了保留和增强，而较为繁杂的背景信息则相对被削弱了一些，从而提供了帧序列的语义信息。

2.2　基于 DBMGU 与 Haar 特征预处理的视频转文字方法

将 DBLSTM 模型应用于视频转文字，在获得效果提升的同时，还应注意到：如前文述及，由于模型参数的明显增多，计算复杂度增大，往往增加了训练时间，不利于将其应用于实时性要求高的场景。针对这些问题，利用 MGU[10]（minimal gated unit）模型的精简性以简化计算参数，从而减少训练时间。而针对视频帧提取 Haar 特征，往往能够提供较好的语义信息，因此，提出基于 DBMGU 模型与特征融合的视频转自然语言方法。

最少门单元 MGU[10]作为 RNNs 模型的一种简化模型，其主要特点是仅具有 1 个门结构，因而被命名为“最少门单元”。在 t 时刻 MGU 单元的计算公式如下（计算符号含义与式(3)~(8)一致）：

$$f_t = \sigma(W_{hf}h_{t-1} + W_{xf}x_t + b_f) \tag{12}$$

$$\tilde{h}_t = \varphi(W_{hh}(f_t \odot h_{t-1}) + W_{xh}x_t + b_h) \tag{13}$$

$$h_t = (1 - f_t) \odot h_{t-1} + f_t \odot \tilde{h}_t \tag{14}$$

MGU 模型的参数远远比 LSTM 要少（相同条件下约为 LSTM 模型的二分之一），理论上其计算复杂度明显低于 LSTM，从而 MGU 模型有效地降低了计算开销，进而提高了训练速度。其次，Chung 等人的研究表明，拥有门结构的 RNNs 类网络，相对于简单地使用双曲正切函数且没有门结构的 RNNs 网络，一般来说，在实验效果上有较显著的提升[6]。MGU 模型遵循了这个结论，保留了必要的一个门结构，使得序列数据的学习效果可以得到保证。为了直观表现 LSTM、MGU 一个时间步内计算复杂度的差异，参考周国兵等人[10]的研究工作，绘出两者的单元结构如图 4 所示。

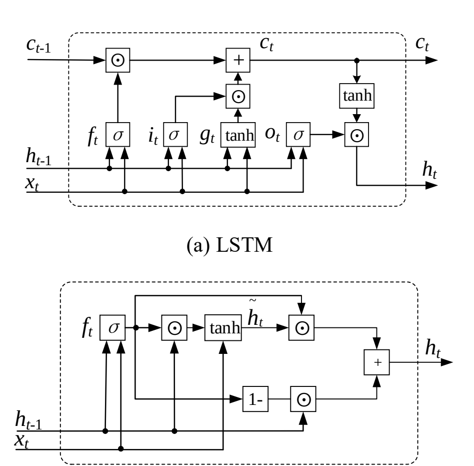


图 4　一个时间步内 LSTM 与 MGU 的计算流程

Fig.4　Calculating process of LSTM and MGU within a time step

DBMGU 的构造方法与 DBLSTM 的构造方法相同。基于 DBMGU 模型和特征融合的视频转文字方法，是将 DBLSTM 模型替换为 DBMGU 模型，同时为了便于对比，也采用经 Haar 特征预处理后的特征来提高模型的效果。

3　模型实验与分析

3.1　实验方法

如图 3 所示,经过 Haar 特征预处理后，视频中的主体目标得到了增强，而其他次要目标则相对地被削弱了，从而提供了一定的语义信息，再对这些视频帧提取 CNNs 特征，因而获得包含了 Haar 特征的 CNNs 特征，这是与原视频帧的 CNNs 特征不同的新特征。同时，鉴于有效的特征融合往往能够提高视频转自然语言描述的准确率和语言效果[11]，本文也将原始视频帧的 CNNs 特征与包含 Haar 特征的 CNNs 特征进行融合，从而增加了训练特征的种类,进而增强了视频特征学习的丰富性，优化学习的效果。如图 5 所示，本文将原始视频帧的 CNNs 特征与包含了 Haar 特征的 CNNs 特征进行融合，以提高视频转自然语言描述的实验效果。

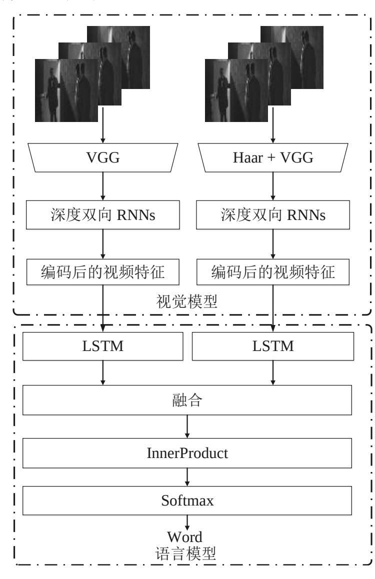


图 5　视频转自然语言实验的模型框架

Fig.5　Model framework of video-to-natural language

在基于小波与 DBLSTM 的视频转自然语言方法中，图中“深度双向 RNN”具体为 DBLSTM 模型。在 DBMGU 与特征融合的视频转自然语言方法中，图中“深度双向 RNN”则具体为 DBMGU 模型。为了便于描述，在以下实验结果图表中，将基于 DBLSTM 和特征融合的视频转自然语言方法简称为“DBLSTM_VGG&Wavelet_Fusion”，将基于 DBMGU 与特征融合的视频转自然语言方法简称为“DBMGU_VGG & Wavelet_Fusion”。

本文使用 Caffe[12]（convolutional architecture for fast feature embedding）深度学习框架实现实验模型。使用 M-VAD 和 MPII-MD 两种主流视频标注数据集独立地对模型进行训练和测试。为了充分利用样本的信息，在训练时综合考虑每个样本的帧序列和词序列的长度，自适应地抽取数量合适的视频帧；在测试时则仅考虑每个样本的帧序列长度对视频帧进行采样。另外，关于视频标注句子与本文方法所生成句子的对比及评价，则使用 METEOR 评测指标作为本文方法输出语句的客观评价指标。METEOR[13]是 Lavir 等人发现在评价指标中召回率的意义后于 2004 年提出的评价指标。他们的研究表明，在考虑了召回率的指标相比于单纯基于准确率的指标，其结果和人工判断的结果有较高相关性。因而 METEOR 评测指标常作为机器翻译、图像转自然语言、视频转自然语言等领域的评价参考，例如 Rohrbach 等人[14]在他们的视频转文字研究中就以 METEOR 作为客观评价指标。模型训练的主要超参数（Hyper Parameter）见表 1。另外，采用的初始学习率为 0.01，学习率调整方法为：每 2 万次迭代将学习率降低为原来的二分之一；训练优化方法为 Mini-Batch 下的 SGD，momentum 设定为 0.9；正则化方法为 Dropout。

表 1　模型训练的主要超参数

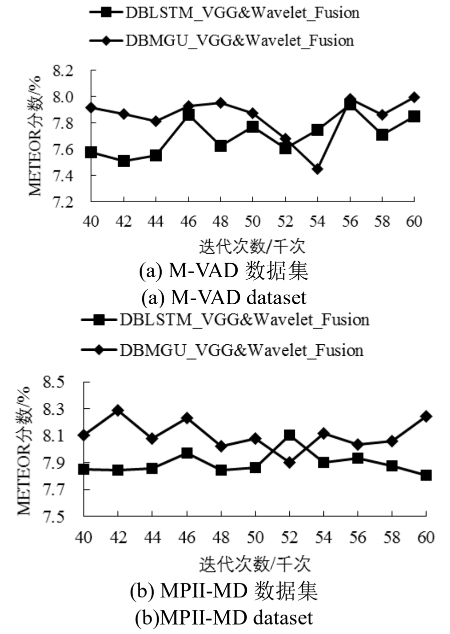
Table 1　Hyper parameter of model training

超参数	值
RNN 模型的 time step	80
RNN 模型的输出向量长度	1000
batch size	16
迭代次数	60000
视频特征降维的全连接层尺寸	4096*500
词向量降维的全连接层尺寸	46168*500
生成词向量的全连接层尺寸	1000*46168

3.2　实验结果分析

参考 Rohrbach 等人[15]的实验分析方法，通过 METEOR 指标评测，以及对标注句子与两种方法生成的句子进行分析比较两个角度，对模型效果进行评估。

首先，在 4 万~6 万次迭代之间，取偶数千次迭代下的模型进行评测，整理得到两种方法的 METEOR 评测分数与迭代次数的关系如图 6 所示。



(a) M-VAD 数据集
(a) M-VAD dataset

(b) MPII-MD 数据集
(b)MPII-MD dataset

图 6　两种方法的 METEOR 评测分数与迭代次数的关系

Fig.6　Relationship between METEOR evaluation scores and iterations of two methods

分析图 6 可知，两种方法的 METEOR 评测分数均较为稳定，可以说明它们在不同的数据集下均能较好地收敛。将图 6

中两种方法的 METEOR 评测分数的峰值整理，并与其他视频转文字方法的 METEOR 评测分数的峰值比较，结果如表 2 所示。METEOR 单位为%，越高效果越佳。

表 2 M-VAD 与 MPII-MD 数据集的 METEOR 评测

Table 2 METEOR Evaluation of M-VAD andMPII-MD datasets

方法	M-VAD	MPII-MD
Visual-Labels[15]	6.3	7.0
Mean pool(VGG)[16]	6.1	6.7
S2VT:RGB(VGG)[2]	6.7	7.1
DBLSTM_VGG&Wavelet_Fusion	7.9	8.1
DBMGU_VGG&Wavelet_Fusion	8.0	8.3

分析表 2 可知，对于 M-VAD 数据集而言，文中提出的 2 种方法在 METEOR 分数上将原 S2VT 模型的 6.7%，分别提高到了 7.9%和 8.0%；同样地，在 MPII-MD 数据集上，文中的 2 个方法，将原 S2VT 模型的 7.1%，分别提高到了 8.1%和 8.3%。一方面，两种方法的 METEOR 评测分数均高于之前的视频转文字方法的 METEOR 分数，这说明深度双向模与特征融合的有机结合，可以提升视频转文字的准确率和语言效果。另一方面，本文提出的两种方法，在视频特征建模部分分别使用的是 DBLSTM 模型和 DBMGU 模型。使用 DBMGU 模型，相对 DBLSTM 模型，METEOR 分数不仅没有降低甚至还略有提高，这说明 DBMGU 模型虽然比 DBLSTM 模型少了近一半的参数，但是所生成的句子与 DBLSTM 所生成的句子在语言效果上相近。更为重要的是，DBMGU 模型可以有效的降低计算复杂度，降低计算开销从而提高视频转文字的速度。

本文对标注句子与两种方法生成的句子进行分析比较。因篇幅限制，每个数据集挑选三个例子作为示例，如图 7 所示。可知，两种方法生成的句子，不仅描述准确，而且相对标注句子包含了更多的细节信息，增加了语言的丰富性。验证了本文所提方法，可以有效提升视频转文字的准确率和语言效果。




手工标注句子： They tour the campus.

DBLSTM_VGG&Wavelet_Fusion： Someone and someone walk up to the sidewalk.

DBMGU_VGG&Wavelet_Fusion： Someone and someone watch as someone and someone walk up to the front lawn.

(a) M-VAD 实例 1

(a) M-VAD example1



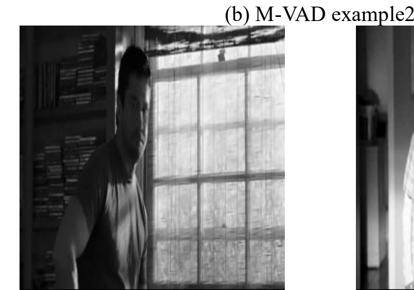

手工标注句子： Now a lunch date.

DBLSTM_VGG&Wavelet_Fusion： Someone glances at someone who smiles and nods.

DBMGU_VGG&Wavelet_Fusion： Someone glances at someone who sits on the couch and smiles at someone.

(b) M-VAD 实例 2

(b) M-VAD example2




手工标注句子： A boy enters.

DBLSTM_VGG&Wavelet_Fusion： Someone turns away and someone follows him to the door.

DBMGU_VGG&Wavelet_Fusion： Someone steps into the living room and finds someone staring at the door then faces someone.

(c) M-VAD 实例 3

(c) M-VAD example3




手工标注句子： She nods sleepily.

DBLSTM_VGG&Wavelet_Fusion： Someone is lying on the bed.

DBMGU_VGG&Wavelet_Fusion： Someone sits on the bed and looks at her.

(d) MPII-MD 实例 1

(d) MPII-MD example1




手工标注句子： He sits up.

DBLSTM_VGG&Wavelet_Fusion： Someone is sitting on the bed.

DBMGU_VGG&Wavelet_Fusion： Someone sits on the bed and looks at the ceiling.

(e) MPII-MD 实例 2

(e) MPII-MD example2




手工标注句子： They turn and walk away together.

DBLSTM_VGG&Wavelet_Fusion： Someone is walking along the sidewalk.

DBMGU_VGG&Wavelet_Fusion： Someone walks up to the front of the house and looks at someone.

(f) MPII-MD 实例 3

(f) MPII-MD example3

图 7 来自 M-VAD 和 MPII-MD 数据集的视频描述实例

Fig.7 Video description examples from M-VAD andMPII-MD datasets

4 结束语

本文针对 S2VT 方法中存在的描述准确率不高的问题，在

DBLSTM 与特征融合的视频转自然语言方法的基础上，提出基于 DBMGU 与特征融合的视频转自然语言方法。所提方法有效地改善了原 S2VT 模型的准确率和语言效果。其中，DBMGU 模型的参数数量仅约为 DBLSTM 模型一半，减少了计算开销，提高了计算速度，却取得了与 DBLSTM 模型相近的语言描述效果，使得所提方法具有广泛的应用场景。当然，当前的工作还存在一些不足，在后期的研究工作中，将针对 S2VT 方法的解码模型、语言模型等方面，做进一步的改进工作。

参考文献：

[1] Kojima A, Izumi M, Tamura T, *et al.* Generating natural language description of human behavior from video images [C]// Proc of the 15th International Conference on Pattern Recognition. Washington DC: IEEE Computer Science, 2000: 728-731.

[2] Venugopalan S, Rohrbach M, Donahue J, *et al.* Sequence to sequence-video to text [C]// Proc of IEEE International Conference on Computer Vision. Piscataway, NJ: IEEE Press, 2015: 4534-4542.

[3] Donahue J, Hendricks L A, Rohrbach M, *et al.* Long-term recurrent convolutional networks for visual recognition and description [J]. IEEE Trans on Pattern Analysis and Machine Intelligence. 2017, 39 (4): 677-691.

[4] Jozefowicz R, Zaremba W, Sutskever I. An empirical exploration of recurrent network architectures [C]// Proc of the 32nd International Conference on Machine Learning. New York: ACM Press, 2015: 2342-2350.

[5] Hochreiter S, Schmidhuber J. Long short-term memory [J]. Neural Computation. 1997, 9 (8): 1735-1780.

[6] Chung J, Gulcehre C, Cho K H, *et al.* Empirical evaluation of gated recurrent neural networks on sequence modeling [EB/OL]. (2015) [2018-06-01]. https://arxiv. org/abs/1412. 3555.

[7] Graves A, Jaitly N, Mohamed A R. Hybrid speech recognition with Deep Bidirectional LSTM [C]// Proc of IEEE Workshop on Automatic Speech Recognition and Understanding. Piscataway, NJ: IEEE Press, 2013: 273-278.

[8] Papageorgiou C P, Oren M, Poggio T. A general framework for object detection [C]// Proc of IEEE International Conference on Computer Vision. Piscataway, NJ, IEEE Press, 2002: 555-562.

[9] Simonyan K, Zisserman A. Very deep convolutional networks for large-Scale image recognition [EB/OL]. (2014) [2018-06-01]. https://arxiv. org/abs/1409. 1556.

[10] Zhou Guobing, Wu Jianxin, Zhang Chenlin, *et al.* Minimal gated unit for recurrent neural networks [J]. International Journal of Automation and Computing. 2016, 13 (3): 226-234.

[11] 梁锐，朱清新，廖淑娇，等. 基于多特征融合的深度视频自然语言描述方法 [J]. 计算机应用. 2017, 37 (4): 1179-1184. (Liang Rui, Zhu Qingxin, Liao Shujiao, *et al.* Deep natural language description method for video based on multi-feature fusion [J]. Journal of Computer Applications. 2017, 37 (4): 1179-1184.)

[12] Jia Yangqing, Shelhamer E, Donahue J, *et al.* Caffe: convolutional architecture for fast feature embedding [C]// Proc of the 22nd ACM International Conference on Multimedia. New York: ACM Press, 2014: 675-678.

[13] Denkowski M, Lavie A. Meteor universal: language specific translation evaluation for any target language [C]// Proc of the 9th Workshop on Statistical Machine Translation. Cambridge, MA: MIT Press, 2014: 376-380.

[14] Rohrbach A, Rohrbach M, Schiele B. The long-short story of movie description [C]// Proc of German Conference on Pattern Recognition. Berlin: Springer, 2015: 209-221.

[15] Rohrbach A, Rohrbach M, Tandon N, *et al.* A dataset for movie description [C]// Proc of IEEE Conference on Computer Vision and Pattern Recognition. New York: IEEE Press, 2015: 3202-3212.

[16] Venugopalan S, Xu Huijuan, Donahue J, *et al.* Translating videos to natural language using deep recurrent neural networks [C]// Proc of Annual Conference of the North American Chapter of the ACL. Cambridge, MA: MIT Press, 2015: 1494-1504.